ChatPub : GPT4-based Web Application to help finding Policy for Korean Youth

Kangsan Kim(rkdtsk419@g.skku.edu), Seungbin Yang(didtmdqlsdl@gmail.com)
Jinho Park(jinho142857@gmail.com), Changmin Jun(jsjs1209@g.skku.edu)

Sungkyunkwan University
2066, Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do, Republic of Korea

October 5, 2023

Abstract

In this report, we will demonstrate a new approach combining Chat-GPT and RAG. Our goal is to increase the accessibility of youth policies to people in Korea. This access will allow young people to more easily obtain policy information, determine eligibility, and apply quickly. As a demo, we created a web application that receives the user's personal information and recommends policies. These attempts will increase the accessibility of youth policies and provide many opportunities to young people who do not receive various benefits. Additionally, in an environment limited to the metropolitan and Gyeonggi areas, it can be applied to many policies across the country and around the world.

Keywords— A I, Chat-GPT, Youth Policy, Retrieval Augmented Generatio

1 Introduction

A substantial portion of the budget is allocated to welfare and housing programs for young individuals, and this budget is gradually increasing. However, a significant challenge lies in the fact that many people are either unaware of the existence of youth policies or, even if they are aware, struggle to identify the policy that best aligns with their needs, consequently missing out on vital benefits. These challenges are compounded by the dispersion of policy information across multiple websites, making it arduous to access comprehensive and pertinent details. The barriers to accessing these policies often appear insurmountable.

Recognizing the pressing necessity to bridge this information gap and ensure broader access to entitled benefits, there is a critical demand for services that can streamline and present youth-oriented policies in a user-friendly manner. Moreover, the potential to dramatically simplify the process arises if government-supported policies, tailored to individual circumstances, can be intelligently recommended based on personal information. This convergence of challenges and opportunities has catalyzed the search for innovative solutions.

In the era of advancing technology, we are witnessing the emergence of various chatbots employing Chat-GPT, offering a promising avenue for improved service delivery. Within this context, the concept of harnessing the capabilities of these chatbots to provide more accessible and personalized government support services has taken shape.

Furthermore, in addition to the chatbot function, our plans extend to various domains, including education, cultural exchange, leisure activities, and, of course, youth policy. In summary, our vision entails creating a community for the younger generation and presenting a unique platform tailored to the needs of young people.

2 Motivation/ Objective

2.1 Strengthen accessibility to youth policy information delivery

The primary objective of this study is to enhance the accessibility and efficiency of delivering youth policy information. We aim to provide young individuals with an experience that simplifies the process of policy application, reducing the time spent on policy searching through the Chat-Pub system and ensuring they can access a wide range of information to maximize the benefits available to them.

2.2 Establishing and activating a youth policy community

Another central goal is to foster the creation of a dynamic community where young individuals can effortlessly communicate, share information, and seek support regarding various youth policies. By consolidating a diverse array of policy information into a single accessible platform, this approach aims to render policy content readily visible and available, empowering young people to independently navigate the intricacies of youth policy.

2.3 Streamlining the youth policy application process

Our third objective is to simplify the application process for youth policies by offering recommendations for informative websites that provide guidance on the application procedures. This approach is devised to enhance accessibility and provide young individuals with a means to expeditiously apply for policies, even when confronted with complex and challenging application processes.

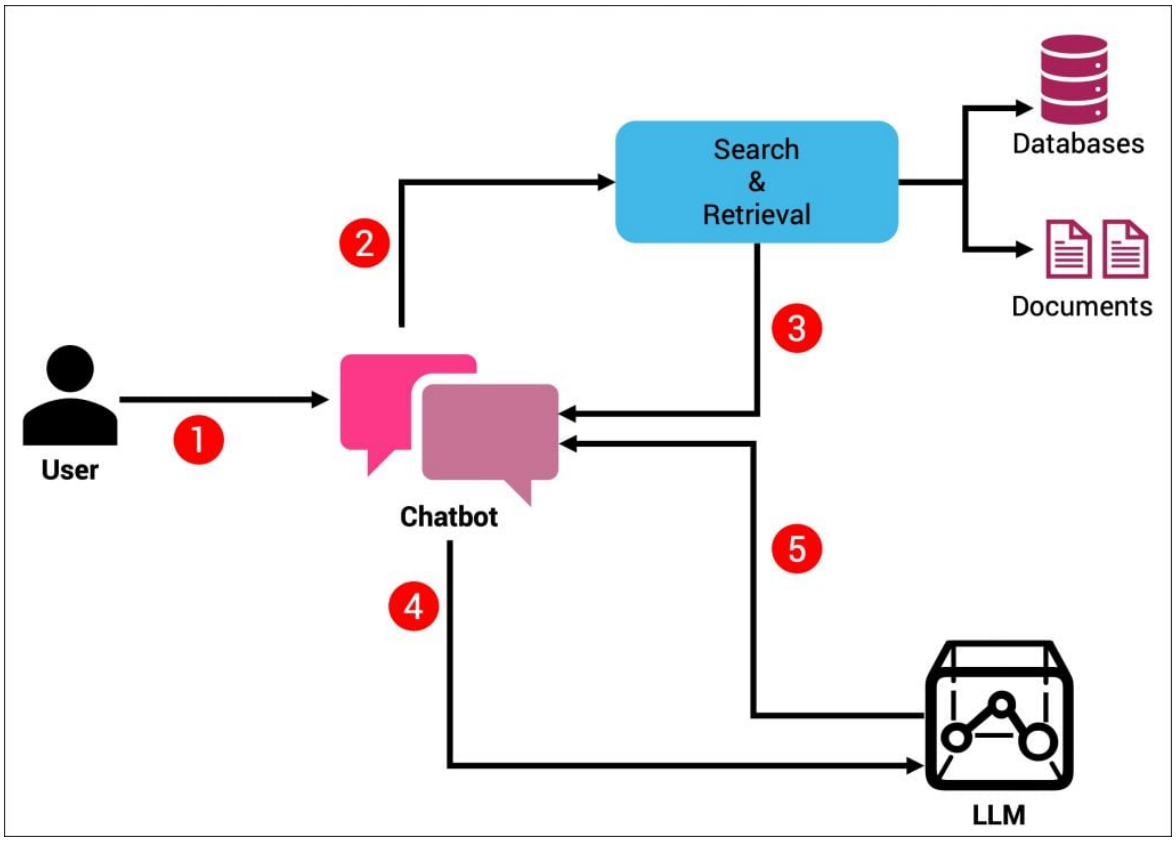

Figure 1: Pipeline[1]

Through the pursuit of these goals, this project will significantly augment the accessibility and efficacy of youth policy, ensuring that young individuals have access to accurate information, engage with supportive communities, and effectively apply for policies that can substantially benefit them.

3 Technical Background

3.1 Text embedding

Text embedding is a technique in natural language processing (NLP) that aims to convert text into numerical representations (e.g. vectors). This transformation enables computers to process and understand textual information. The core principle behind text embeddings is to capture the semantic meanings of words or phrases such that semantically similar items are mapped to proximate points in the embedding space. Popular methods like Word2Vec and BERT have been foundational in achieving this representation. The embedded vectors have been widely utilized in numerous NLP tasks, including sentiment analysis and text classification, due to their ability to capture contextual and semantic information in the text.

3.2 Sentence Transformer

Sentence Transformer is a Python framework for state-of-the-art sentence, text, and image embeddings. While traditional embeddings like Word2Vec focus on individual words, Sentence Transformers aim to capture longer text sequences' meaning and context. Built on top of transformer architectures such as BERT, these models are trained so that similar sentences are close in the vector space, preserving semantic meanings across sentences. This representation becomes especially beneficial in tasks like semantic textual similarity and information retrieval, where understanding the broader context of text sequences is essential.

3.3 Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is an NLP technique that intertwines the robustness of pre-trained language models with the capability to retrieve and utilize external knowledge. The mechanism comprises knowledge retrieval and sequence generation steps, typically using a transformer-based architecture like GPT. During the knowledge retrieval step, relevant documents or text snippets are initially fetched from an external knowledge source (such as a corpus of text) based on the query or input. Subsequently, the retrieved context is used to inform the generation of a response or completion in the second step. RAG has shown remarkable efficacy in various applications like open-domain question answering, where the model can retrieve relevant knowledge, and dialogue systems where maintaining context and providing informative replies is pivotal. The ability of RAG to harness external knowledge enables it to generate responses that are not only contextually relevant but also rich in factual content.

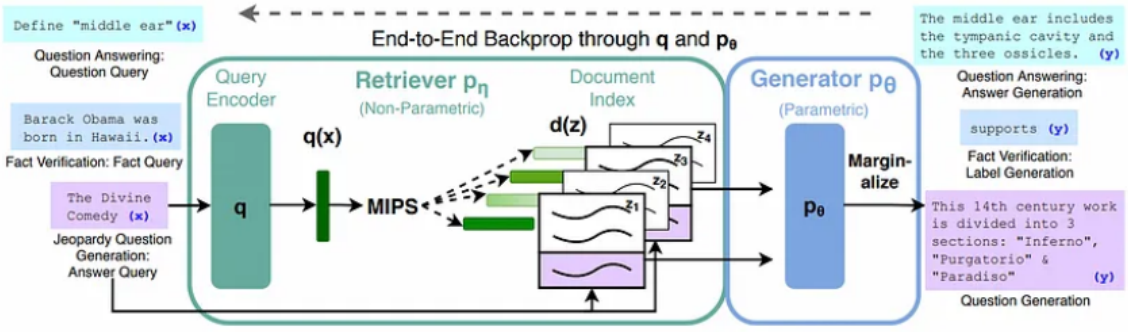

Figure 2: RAG overview[2]

4 Problem Statement/ Proposed Solution

4.1 Problem Statement

4.1.1 Low Accessibility

Despite the existence of numerous policies for young people, there is a lack of accessibility to youth policies. Sometimes, young people may not receive enough information about policies, or the communication methods used may be complex, making it difficult for them to understand and utilize the policies. If young people are not directly involved in policy formulation, the resulting policies may not align with their needs and realities. This lack of involvement can lead to a lack of interest and engagement. To address these issues, effective information dissemination, we introduced a chatbot service.

4.1.2 The Unnecessary Manpower Problem

There is unnecessary manpower being consumed on policy websites. Replacing the manpower involved in the operation of the young policy website with a chatbot can handle this problem. Chatbot can interact with users in real-time 24/7 through automated response systems. By quickly responding to user queries and providing necessary information immediately, the need for manpower in individualized responses can be significantly reduced. Additionally, there is an initial investment in building a chatbot, operational and maintenance costs are generally lower compared to employing personnel for similar tasks on a website.

4.2 Proposed Solution

4.2.1 building a policy database

To construct the embedding space, initially, web scraping of a youth policy website is conducted. The textual data obtained through scraping is transformed into embedding vectors utilizing a sentence transformer. The sentence transformer employs a model subjected to multi-task learning using the KorSTS and KorNLI training datasets.[2] [3] [4] Subsequently, the Faiss library is utilized to store the vectors in the embedding space.

4.2.2 Customized Youth Policy Chatbot

We will construct a chatbot capable of providing pertinent policy information by utilizing user's basic personal data and inquiries. The chatbot adheres to the RAG structure, which can be distinctly categorized into two main components: the knowledge retrieval and generation parts. In the knowledge retrieval part, policy information necessary for generating responses is extracted from a pre-constructed embedding space using the user's personal data and inquiries. We convert the user's personal data and inquiries into vectors utilizing the same sentence transformer model employed to construct the embedding space. The transformed vectors undergo cosine similarity computations with the vectors residing in the embedding space, following which a subset of information snippets with high similarity are extracted. In the generation part, user-tailored policy responses are produced using the extracted information snippets. The user's query and the extracted information snippets are amalgamated, serving as inputs for the generative model. The model is furnished to respond to the user's query by referring to the provided information snippets. For instance, if a user's query is 'Tell me about the employment policy.' and the extracted information snippets pertain to the 'Seoul Youth Happiness Employment Policy.', the input for the generative model would be framed as: 'You are required to respond to the user's query utilizing the provided context. Context: Seoul Youth Happiness Employment Policy. User Query: Tell me about the employment policy.' In this research, we employ the GPT-4 model, provided by OpenAI, as our generative model.

5 Planning in detail

5.1 Task assignment

The project is largely divided into two tasks: AI and Web development. The work was decided based on the interests and capabilities of each team member. AI is a model and data collection, and web development is divided into front and back.In addition to the designated tasks of the members, we will flexibly divide the additional tasks as needed.

Name	Task
Kangsan Kim	Team Leader, UI/UX
Seungbin Yang	AI Engineer
Jinho Park	Data Engineer
Changmin Jun	Back-End Developer

Table 1: Task table.

5.2 Project schedule

Based on the tasks that we divided above, we set the project schedule briefly. The schedule can be modified depending on the situation.

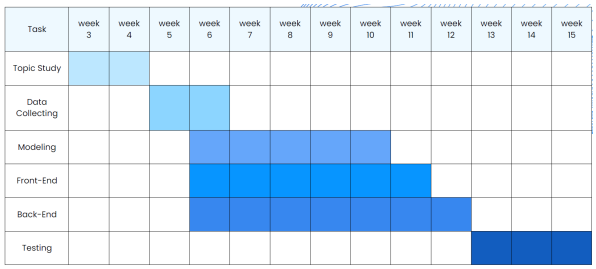

Figure 3: Project schedule

References

[1] Janakiram MSV. Freshen up llms with retrieval augmented generation. *https://thenewstack.io/freshen-up-llms-with-retrieval-augmented-generation/*, 2023.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[4] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*, 2020.